Rede INTERAC, 10 de Julho de 2008, UM

Integrating Multimedia Information Retrieval, Signal Processing and Human Computer Interaction

LASIGE-HCIM: Human-Computer Interaction and Multimedia Research Team

Carlos Teixeira

This abstract presents current research goals of the author, most of which were built within the framework of the author's research group HCIM (Human-Computer Interaction and Multimedia Research Team at LASIGE-DI/FCUL). This group had long term projects in the area of digital talking books and hypermedia (Chambel and Guimarães, 2002). This was an inspiration for the creation of new multimedia contents and tools intertwining audio-visual and textual related material, linking different expression modes coherently in a "multimedia object". Adding higher level information that cuts across the several modalities (text, audio, images and film) significantly augment the range of functionalities to be considered. Higher levels of information will then be used to provide high level summaries or multimedia indexes of the (mixed) content, drawing from work in ontology extraction from text, text classification and natural language generation techniques (Lawrie and Croft, 2003). Using relationships between text (near or aligned) and pictures or video-clip can increase connectivity between the different modalities and provide truly multimedia abstracts. Research is conducted on the subjects of *auto-illustrate* (link images or videos to text) and *auto-annotate* (link text to audiovisual objects, Barnard *et al.* 2003).

Drawing from work in areas as diverse as computational stylistics, ontology validation and terminological systems, it is now possible to assign a set of labels to a text, covering widely different dimensions:

- text structure (introduction, action, commentary, dialogue, different scenes in a fiction book, or in a carefully edited TV program, etc.)
- events (roles, date, place, causes and consequences)
- topic, subject, news cluster, named entities, geographical scope, etc., and
- establishing links between subjects, events and texts

The goal here is to go one step further than systems which allow people to access individual scenes, such as MUSCLE's example of "give me all video clips where JK is to the right of his assistant" (Enis Cetin, 2005) to more abstract (and arguably more useful) notions such as "gun firings in Iraq" (news as raw material) or "TV debates where the consequences of the oil raising price are named" or "pictures of Algarve this Summer" for other kinds of plausibly relevant information access for different kinds of users.

Several applications are envisaged for providing integrated multimedia browsing, querying and production of the new contents. In a first phase, partially annotated related multimedia contents are selected in order to allow the construction of new integrated contents. This requires the development of specialized text aligning algorithms. A new text alignment was presented in (Teixeira and Respício, 2007), not based on the classical longest subsequence approach, with new features supporting new applications. Further advanced techniques such as automated knowledge discovery and extraction, topic detection and automatic summarization, are expected to improve this applications.

Browsing the new integrated contents will employ summarization capabilities (Tsoneva, *et al.* 2007) in order to provide multimedia abstracts at several levels of detail, as well as following different knowledge organization strategies. This is one the anchor points where collaboration with other more specialized research teams can be envisaged. In a second phase, larger non-annotated raw multimedia data is expected to be automatically processed. This opens collaboration with teams working in signal processing for audio, image and video.

The applications envisaged above should provide multimedia results that can be efficiently perceived and sometimes also changed by humans. Further on, the expected production tools, which are designed to automatically integrate relatively large pieces of multimedia contents, can mostly be considered as preprocessing tools. In order to produce final quality contents, multimedia content producers demands additional facilities for reviewing/manual post-edition of every automatic derived result. For this, as well as for browsing, and querying the final content, it is crucial to design efficient usable human interfaces. For the production phases these interfaces should enhance the detection of crucial errors or incoherencies in the integrated contents. On the other hand, for the content end users, interfaces should enhance features that can provide better usability for browsing and querying (Hammoud and Mohr, 2000).
End-user consumption of audio-visual content will be greatly enhanced with support for mixed-content queries, where users are able to provide query terms both in audio and through a text input form field. These multimodal capabilities together with context awareness allowed by the indexed and annotated media will offer to end-users a more natural and effortless interaction experience, while increasing the usage scenarios.

“To get closer to the vision of useful multimedia-based search and retrieval, the annotation and search technologies need to be efficient and use semantic concepts that are natural to the user. This requires tagging and/or annotation of multimedia content with semantic concepts describing digital objects and in more appealing applications using descriptions of emotions and related human expressions. However, semantic annotation mostly depends on human interaction, which is expensive, time consuming and therefore infeasible for many applications. Even the annotation of few hundreds of personal images captured during a single year by a single person is a tedious task that nobody wants to do. As a consequence, multimedia structuring, annotation and retrieval using semantic structures and descriptions of emotions which are natural to humans remains critical” (ICIP-MIR, 2008).

Finally, some work will have to be done with this new multimedia objects concerning the use of standards and digital rights.

LASIGE is located in the Faculty of Sciences from the University of Lisbon. This University provides a unique campus in Lisbon area where multidisciplinary synergies can be found for the anchor areas described above, namely those concerning signal and text processing, as well as many domains of information systems (considering for instance the both well known Faculty of Medicine and the Faculty of Letters). Following this in-house perspective, some work concerning underwater acoustic signals for fish communication was already started with the Department of Animal Biology of the Faculty of Sciences (Amorim *et al.*, 2008).

References

Amorim, M. C. P., Simões, J. M. and Fonseca, P. J., 2008. "Acoustic communication in the Lusitanian toadfish, Halobatrachus didactylus: evidence for an unusual large vocal repertoire". Journal of Marine Biological Association of the UK, (in press) doi:10.1017/S0025315408001677

Barnard, K., Duygulu, P., Freitas, N., Forsyth, D., Blei, D., Jordan, M., 2003. "Matching words and pictures". Journal of Machine Learning Research **3**, pp.1107-1135.

Chambel, T. and Guimarães, N., 2002. "Context Perception in Video-Based Hypermedia Spaces", Proceedings of ACM Hypertext'02, College Park, Maryland, USA.

Enis Cetin, A., 2005. "Report on progress with respect to partial solutions on human detection algorithms, human activity analysis methods, and multimedia databases". Deliverable 11-4, http://www.cs.bilkent.edu.tr/~ismaila/MUSCLE/DeliverableWP11_4.doc.

Hammoud, R., and Mohr, R., 2000. "Interactive Tools for Constructing and Browsing Structures for Movie Films". Proceedings of the eighth ACM International Conference on Multimedia. Marina del Rey, California, 497–498.

ICIP-MIR, 2008. First ICIP Workshop on Multimedia Information Retrieval: "New Trends and Challenges", San Diego, California, http://icip08-mir.qmul.net/.

Lawrie, D. J. and Croft, W. B., 2003. "Generating hierarchical summaries for web searches." In Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in informaion Retrieval (Toronto, Canada, July 28 - August 01, 2003). SIGIR '03. ACM, New York, NY, 457-458. DOI= http://doi.acm.org/10.1145/860435.860549

Lienhart, R., Pfeiffer, S., and Effelsberg, W., 1997. "Video Abstracting", Communications of the ACM, Vol. 40, No. 12, 55-63.

Teixeira, C. and Respício, A., 2007. "See, Hear or Read the Film". in Ma L., Rauterberg M. and Nakatsu R. (eds.): Entertainment Computing - ICEC2007, Proceedings of the 6th International Conference on Entertainment Computing, Lecture Notes in Computer Science, 4740, Springer, 271-281.

Tsoneva, T., Barbieri, M, and Weda, H., 2007. "Automated Summarization of Narrative Video on a Semantic Level". International Conference on Semantic Computing. ICSC2007. IEEE Press, 169 – 176.